Attorney Docket Number 2102299-991130

#4

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: Mandyam et al.

Serial No.: 10/056,300

Filed: January 22, 2002

Group Art Unit: 2176

Examiner: Not Yet Assigned

Title: **METHOD FOR EXTRACTING CONTENT FROM STRUCTURED OR UNSTRUCTURED TEXT DOCUMENTS**

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as FIRST CLASS MAIL in an envelope addressed to: Assistant Commissioner for Patents, Box Missing Parts, Washington, D.C. 20231, on

April 25, 2002

Yvette Renee Killingham

* * *

PRELIMINARY AMENDMENT

Assistant Commissioner for Patents
Box Missing Parts
Washington, D.C. 20231

Sir/Madam:

Please enter the following preliminary amendments for the above-referenced application:

IN THE SPECIFICATION:

Please amend the following paragraphs of the specification (marked-up versions of the following amended paragraphs are attached hereto as Appendix A):

Replacement paragraph at page 3, lines 7 – 10 (clean version):

One non-limiting advantage of the invention is that it presents a method for defining selection commands for both structured and unstructured documents. Structured documents can be interpreted as having structural content and textual/character content. Unstructured documents can only be interpreted as having textual/character content.

Replacement paragraph at page 14, line 22 – page 15, line 2 (clean version):

As shown in Figure 4, a selection envelope 1400 is a container for a section of a source document 1100, delineated by two markers referred to as the *begin marker* 1200 and *end marker* 1300. These markers are virtual delineators that are created only during runtime by selection command 1600. The begin marker 1200 defines the beginning of the selection envelope 1400 while the end marker 1300 defines the end of the selection envelope. The selected contents 1500 is what lies between these two markers.

Replacement paragraph at page 16, lines 1 – 8 (clean version):

For structured documents, a selection envelope can contain various arrangements of structures. As shown in Figure 5, a structured document may be represented as a hierarchical structure 1110, including a parent object 1111, child objects 1112, 1114, and descendants 1113, 1115. A selection envelope 1410 made of a begin marker 1210 and end marker 1310 may contain any valid structural element represented by object 1112 and descendants 1113. Selection envelopes containing structural objects place their begin markers and end markers immediate adjacent to the object so that they exclusively define the desired object. Just as the structure of a document may exist as an abstract system created by an XML processor, the begin and end markers are virtual objects in the document.

Replacement paragraph at page 16, lines 9 – 12 (clean version):

For unstructured documents, a selection envelope can contain contiguous segments of text based on the textual representation of the document. An example of a selection envelope 1420 with relation to an unstructured document 1120, which begins at location 1121 and ends at location 1122, is shown in Figure 6. Begin marker 1220 and end marker 1320 are positioned around segments of content 1520 within the document 1120, near possible locating strings 1130, 1131, respectively.

Replacement paragraph at page 16, lines 13 – 18 (clean version):

More generally, a system of selection envelopes can be defined so that each successive selection envelope, or child envelope, is defined relative to a previously defined envelope, or parent envelope. As shown in Figure 7, selection envelope 1430 may be defined for source document 1100 via selection command 1601. Envelope 1430 may then be used to produce envelope 1431 via selection command 1602, envelope 1431 may the produce envelope 1432 via selection command 1603, and envelope 1432 may produce envelope 1433 via selection command 1604, thereby creating a series of nested selection envelopes 1400 having begin markers 1200 and end markers 1300. Selection commands are more fully explained below.

Replacement paragraph at page 16, lines 19 – 24 (clean version):

The relationship between a parent envelope and its successor, or child envelope can take form in one of three ways. A child selection envelope 1441 may be either nested within a parent selection envelope 1440, as shown in Figure 8; partially overlapping a parent selection envelope 1440, as shown in Figure 9; or completely outside of a parent selection envelope 1440, as shown in Figure 10. The scope of the selection is iteratively refined until the desired content has been selected.

Replacement paragraph at page 16, lines 19 – 24 (clean version):

Furthermore, multiple sets of selection envelopes may exist simultaneously for a given document when a selection command is applied. Referring to Figure 11, a structured document 1110 can be seen to have two selection envelopes 1410 and 1411 (e.g., having begin and end markers 1210, 1310 and 1211, 1311, respectively) that contain two different object structures. Referring to Figure 12, an unstructured document 1125 (beginning at location 1126 and ending at location 1127) can be seen to also have two selection envelopes 1421, 1422, having begin and end markers 1221, 1321 and 1222, 1322, respectively.

Replacement paragraph at page 18, lines 11 – 17 (clean version):

For structured documents 1110, the general relationship between selection commands and selection envelopes is illustrated in Figure 14. A selection command 1610 may identify an object structure composed of a child object 1112 and descendant objects 1113, and thus specify a selection envelope 1410 around the structure. For unstructured documents 1120, this general relationship is illustrated in Figure 15. A selection command 1620 may define the locations of the virtual begin marker 1220 and virtual end marker 1320 and thus, define a selection envelope 1420.

Replacement paragraph at page 23, lines 3 – 7 (clean version):

The method 2000 will be defined as follows for the following four examples. The source document Y is an HTML document, seen in rendered form in Figure 16 and in HTML source view in Figure 17. The examples will illustrate the creation of four selection envelopes s_1, s_2, s_3, and s_4 that respectively identify x_1, x_2, x_3, and x_4. As described above, selection envelopes are functions of selection commands 'c' that are defined below.

Replacement paragraph at page 26, lines 16 – 18 (clean version):

To further elaborate on the use of selection commands, the second table of document Y will be selected for use in Y'. This again illustrates the use of position or sequential index of an object within a parent selection envelope.

Replacement paragraph at page 29, lines 2 – 7 (clean version):

Referring to step 2001, the desired content has not yet been selected thus necessitating the definition of another selection envelope. For this second selection envelope, the source in step 2004 is document Y and $\mathbf{x}_3^1$. For step 2005, selection command c_k^1 has not yet been chosen. To determine c, the process of Figure 13B is again followed. Step 2016 dictates that either structural, pattern-based or any combination of commands c_1, c_2, or c_3 can be used.

Replacement paragraph at page 29, lines 8 – 9 (clean version):

For step 2017, the first selection command is determined to be a structural selection command 2013, as seen in Figure 13B. Command $\mathbf{c}_1$ is parameterized as follows:

Replacement paragraph at page 30, lines 12 – 16 (clean version):

This selection envelope example illustrates the use of a command that combines structural and pattern-based commands. Yet again, the process of Figure 13A is used. Step 2004 defines the source information for envelope specification; in this case, the source is document Y, as shown in Figure 17. For step 2005, a selection command c_k^1 is to be selected from the set of functions C defined above and then parameterized.

Replacement paragraph at page 30, lines 17 – page 31, line 16 (clean version):

In order to do this, steps 2016 and 2017 of the process in Figure 13B are used. Given that document Y is structured, step 2016 of the process seen in Figure 13B allows either structural, pattern-based or any combination of commands c_1, c_2, or c_3 to be used. For the purposes of the example, the desired content x_4, is deemed to be reliably extractable by immediately using a selection command c_3. Command c_3 combines structural and pattern-based commands using programmatic constructs. Thus for step 2017, both a structural/contextual selection command 2013 and a pattern-based selection command 2015 are selected. The selection command $\mathbf{c_3}$ is parameterized as follows:

type = row
instance = 1
string = “Row1”
inclusion = true

Thus, $\mathbf{c_3}$ is such that

c_3 defines a resulting selection envelope, s_4, such that:
$s_4 = f(c_3)$
which is equivalent to equation (5) above. Stated another way,
$s_4 = c_3\ (Y) = x_4$

where $\mathbf{x_4}$ can be seen in Figure 23. As the desired content has been selected, the answer for step 2001 is ‘yes’ and the selected content x_4 is available for use in Y’.

IN THE CLAIMS:

Please amend claim 1 (a marked-up version of claim 1 is attached hereto as Appendix B) and add new claims 2 – 18 as follows:

Claim 1 (amended, clean version)

A method for extracting content from a document, comprising the steps of:

creating at least one selection envelope based upon a plurality of selection commands for locating specific content within said document; and

selecting content from said document based upon said at least one selection envelope.

Claim 2 (new)

The method of claim 1 wherein said selection envelope comprises a begin marker and an end marker, which respectively define the beginning and end of said selection envelope.

Claim 3 (new)

The method of claim 1 wherein said at least one selection envelope comprises a parent envelope and a child envelope.

Claim 4 (new)

The method of claim 3 wherein said child envelope is nested within said parent envelope.

Claim 5 (new)

The method of claim 3 wherein said child envelope partially overlaps said parent envelope.

Claim 6 (new)

The method of claim 3 wherein said child envelope is completely outside of said parent envelope.

Claim 7 (new)

The method of claim 1 wherein said plurality of selection commands comprises a command based on document structure.

Claim 8 (new)

The method of claim 1 wherein said plurality of selection commands includes a command based on a character pattern.

Claim 9 (new)

The method of claim 1 wherein said plurality of selection commands comprises a combined command based on both document structure and a character pattern.

Claim 10 (new)

A method for extracting content from a source comprising the steps of:

identifying said source for extracting content;

parameterizing at least one selection command to operate on said source;

defining a selection envelope to select desired content from said source by use of said at least one selection command;

selecting content from said source by use of said selection envelope;

determining whether said desired content has been selected; and

extracting said selected content if said desired content has been selected.

Claim 11 (new)

The method of claim 10 further comprising the steps of:

defining a second selection envelope by use of at least one second selection command if said desired content has not been selected;

selecting content from said source by use of said second selection envelope;

determining whether said desired content has been selected; and

extracting said selected content if said desired content has been selected.

Claim 12 (new)

The method of claim 11 wherein said first selection envelope comprises a parent envelope and said second selection envelope comprises a child envelope.

Claim 13 (new)

The method of claim 10 wherein said source comprises a document.

Claim 14 (new)

The method of claim 10 wherein said source comprises a section of a document.

Claim 15 (new)

The method of claim 10 wherein said step of parameterizing said at least one selection command includes determining whether said source is structured or unstructured, and selecting said at least one selection command is based upon this determination.

Claim 16 (new)

The method of claim 15 wherein said at least one selection command comprises a structure based command selected from the group including select by name commands, select by location commands, select by sibling relationship commands, select by attribute commands and select by counter commands.

Claim 17 (new)

The method of claim 15 wherein said at least one selection command comprises a character based command selected from the group including select text contain commands and select text matching pattern commands.

Claim 18 (new)

The method of claim 15 wherein said at least one selection command comprises a combined structure and character based command.

It is respectfully asserted that the foregoing amendments place the application in better condition for examination, and that none of the foregoing changes contain any new matter.

The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. 07-1896.

Respectfully submitted,

GRAY CARY WARE & FREIDENRICH LLP

Dated: April 22, 2002

By D. Alberti

David Alberti
Reg. No. 43,465
Attorney for Applicants

GRAY CARY WARE & FREIDENRICH
1755 Embarcadero Road
Palo Alto, California 94303-3340
Telephone: (650) 833-2052
Facsimile: (650) 320-7401

APPENDIX A

Replacement paragraph at page 3, lines 7 – 10 (marked-up version):

One non-limiting advantage of the invention is that it presents a method for defining selection commands for both structured and unstructured documents. Structured documents can be interpreted as having structural content and textual/character content. Unstructured documents can only be interpreted as having [textural]textual/character content.

Replacement paragraph at page 14, line 22 – page 15, line 2 (marked-up version):

As shown in Figure 4, a selection envelope 1400 is a container for a section of a source document 1100, delineated by two markers referred to as the *begin marker* 1200 and *end marker* 1300. These markers are virtual delineators that are created only during runtime by selection command 1600. The begin marker 1200 defines the beginning of the selection envelope 1400 while the end marker 1300 defines the end of the selection envelope. The selected contents 1500 is what lies between these two markers.

Replacement paragraph at page 16, lines 1 – 8 (marked-up version):

For structured documents, a selection envelope can contain various arrangements of structures. As shown in Figure 5, a structured document may be represented as a hierarchical structure 1110, including a parent object 1111, child objects 1112, 1114, and descendants 1113, 1115. A selection envelope 1410 made of a begin marker 1210 and end marker 1310 may contain any valid structural element represented by object 1112 and descendants 1113. Selection envelopes containing structural objects place their begin markers and end markers immediate adjacent to the object so that they exclusively define the desired object. Just as the structure of a document may exist as an abstract system created by an XML processor, the begin and end markers are virtual objects in the document.

Replacement paragraph at page 16, lines 9 – 12 (marked-up version):

For unstructured documents, a selection envelope can contain contiguous segments of text based on the textual representation of the document. An example of a selection envelope 1420 with relation to an unstructured document 1120, which begins at location 1121 and ends at location 1122, is shown in Figure 6. Begin marker 1220 and end marker 1320 are positioned around segments of content 1520 within the document 1120, near possible locating strings 1130, 1131, respectively.

Replacement paragraph at page 16, lines 13 – 18 (marked-up version):

More generally, a system of selection envelopes can be defined so that each successive selection envelope, or child envelope, is defined relative to a previously defined envelope, or parent envelope. As shown in Figure 7, selection envelope 1430 may be defined for source document 1100 via selection command 1601. Envelope 1430 may then be used to produce envelope 1431 via selection command 1602, envelope 1431 may the produce envelope 1432 via selection command 1603, and envelope 1432 may produce envelope 1433 via selection command 1604, thereby creating a series of nested selection envelopes 1400 having begin markers 1200 and end markers 1300. [and so on.] Selection commands are more fully explained below.

Replacement paragraph at page 16, lines 19 – 24 (marked-up version):

The relationship between a parent envelope and its successor, or child envelope can take form in one of three ways. A child selection envelope 1441 may be either nested within a parent selection envelope 1440, as shown in Figure 8; partially overlapping a parent selection envelope 1440, as shown in Figure 9; or completely outside of a parent selection envelope 1440, as shown in Figure 10. The scope of the selection is iteratively refined until the desired content has been selected.

Replacement paragraph at page 16, lines 19 – 24 (marked-up version):

Furthermore, multiple sets of selection envelopes may exist simultaneously for a given document when a selection command is applied. Referring to Figure 11, a structured document 1110 can be seen to have two selection envelopes 1410 and 1411 (e.g., having begin and end markers 1210, 1310 and 1211, 1311, respectively) that contain two different object structures. Referring to Figure 12, an unstructured document 1125 (beginning at location 1126 and ending at location 1127) can be seen to also have two selection envelopes 1421, 1422, having begin and end markers 1221, 1321 and 1222, 1322, respectively.

Replacement paragraph at page 18, lines 11 – 17 (marked-up version):

For structured documents 1110, the general relationship between selection commands and selection envelopes is illustrated in Figure 14. A selection command 1610 may identify an object structure composed of a child object 1112 and descendant objects 1113, and thus specify a selection envelope 1410 around the structure. For unstructured documents 1120, this general relationship is illustrated in Figure 15. A selection command 1620 may define the locations of the virtual begin marker 1220 and virtual end marker 1320 and thus, define a selection envelope 1420.

Replacement paragraph at page 23, lines 3 – 7 (marked-up version):

The method [1000] 2000 will be defined as follows for the following four examples. The source document Y is an HTML document, seen in rendered form in Figure 16 and in HTML source view in Figure 17. The examples will illustrate the creation of four selection envelopes s_1, s_2, s_3, and s_4 that respectively identify x_1, x_2, x_3, and x_4. As described above, selection envelopes are functions of selection commands 'c' that are defined below.

Replacement paragraph at page 26, lines 16 – 18 (marked-up version):

To further elaborate on the use of selection commands, the second table of document Y will be selected for use in Y'. This again illustrates the use of position or sequential index [or] of an object within a parent selection envelope.

Replacement paragraph at page 29, lines 2 – 7 (marked-up version):

Referring to step 2001, the desired content has not yet been selected thus necessitating the definition of another selection envelope. For this second selection envelope, the source in step 2004 is document Y and x_3^1. For step 2005, selection command c_k^1 has not yet been chosen. To determine c, the process of Figure 13B is again followed. Step 2016 dictates that either structural, pattern-based or any combination of commands c_1, c_2, or c_3 can be used.

Replacement paragraph at page 29, lines 8 – 9 (marked-up version):

For step 2017, the first selection command is determined to be a structural selection command 2013, as seen in Figure 13B. Command c_1 is parameterized as follows:

Replacement paragraph at page 30, lines 12 – 16 (marked-up version):

This selection envelope example illustrates the use of a command that combines structural and pattern-based commands. Yet again, the process of Figure 13A is used. Step 2004 defines the source information for envelope specification; in this case, the source is document Y, as shown in Figure 17. For step 2005, a selection command c_k^1 is to be selected from the set of functions C defined above and then parameterized.

Replacement paragraph at page 30, lines 17 – page 31, line 16 (marked-up version):

In order to do this, steps 2016 and 2017 of the process in Figure 13B are used. Given that document Y is structured, step 2016 of the process seen in Figure 13B allows either structural, pattern-based or any combination of commands c_1, c_2, or c_3 to be used. For the purposes of the example, the desired content x_4, is deemed to be reliably extractable by immediately using a selection command c_3. Command c_3 combines structural and pattern-based commands using programmatic constructs. Thus for step 2017, both a structural/contextual selection command 2013 and a pattern-based selection command 2015 are selected. The selection command_$\mathbf{c_3}$ is parameterized as follows:

type = row
instance = 1
string = "Row1"
inclusion = true

Thus, $\mathbf{c_3}$ is such that

c_3 defines a resulting selection envelope, s_4_such that:
$s_4 = f(c_3)$
which is equivalent to equation (5) above. Stated another way,
$s_4 = c_3\ (Y) = x_4$

where $\mathbf{x_4}$ can be seen in Figure 23. As the desired content has been selected, the answer for step 2001 is 'yes' and the selected content x_4 is available for use in Y'.

APPENDIX B

Claim 1 (amended, marked-up version)

A method for extracting content from a document, comprising the steps of:

creating at least one selection envelope based upon a plurality of selection commands for locating specific content within said document; and

selecting content from said document based upon said at least one selection envelope.